ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230589

Task-Sensitive Discriminative Mutual Attention Network for Few-Shot Learning

Baogui Xu[a,b], Chengjin Xu[c], Zhiwu Lu[a,b] and Bing Su[a,b;*]

[a]Gaoling School of Artificial Intelligence, Renmin University of China
[b]Beijing Key Laboratory of Big Data Management and Analysis Methods
[c]International Digital Economy Academy

Abstract. Many few-shot image classification methods focus on learning a fixed feature space from sufficient samples of seen classes that can be readily transferred to unseen classes. For different tasks, the feature space is either kept the same or only adjusted by generating attentions to query samples. However, the discriminative channels and spatial parts for comparing different query and support images in different tasks are usually different. In this paper, we propose a task-sensitive discriminative mutual attention (TDMA) network to produce task-and-sample-specific features. For each task, TDMA first generates a discriminative task embedding that encodes the inter-class separability and within-class scatter, and then employs the task embedding to enhance discriminative channels respective to this task. Given a specific query and different support images, TDMA further incorporates the task embedding and long-range dependencies to locate the discriminative parts in the spatial dimension. Experimental results on miniImageNet, tieredImageNet and FC100 datasets show the effectiveness of the proposed model.

1 Introduction

The success of deep learning depends heavily on a large amount of labeled training data, limiting the scalability of deep learning models to new or rare concepts with few annotations. As a promising approach to tackle this challenge, few-shot classification aims at learning a classifier trained on seen classes that can be adapted to new unseen classes, given only very few labeled samples of new classes.

Most existing few-shot classification methods [36, 32, 9, 25, 18, 28, 8] employ a meta-learning paradigm to discover transferable meta-knowledge from a set of tasks that mimic the few-shot test setting with the seen classes. Different methods differ in the meta-knowledge, which could be a distance metric [36, 32], an optimizator [18, 28], and initial parameters [9, 25]. Meta-knowledge is generally applied to all few-shot classification tasks with unseen classes indiscriminately. As the embedding function is performed on support and query samples independently, the channel-spatial feature map of a sample remains the same compared with any other sample.

However, the discriminative channels and foreground areas of interest are often different when comparing different sample pairs in different tasks. As illustrated in Figur 1, in three different tasks and when comparing with different images, the discriminative parts of the same support image differ a lot. Adaptively identifying the most discriminative features and the most relevant regions for each sample pair is important for fully leveraging useful information from few labeled samples. This can be difficult because whether a pixel or a local region is discriminative may depend on the whole task and multiple non-local regions of the compared image. For example, the discriminative parts between rabbit images and cat images might be their ears while the discriminative parts between rabbit images and dog images could possibly be their noses. As far as we know, few work [12] is devoted to adapting the features according to specific sample pairs but only produces spatial cross-attentions with local operations.

Figure 1: Illustration of the main idea for TDMA. Each column shows the support images for a 3-way 1-shot task. The first class of the support images in three different tasks are "bird" class. The discriminative parts of the bird image vary considerably in comparison with other support images in the three tasks. In the first task, the bird's differentiation from the tiger and bear is located in its body, where the bird possesses wings while the others do not. In the second task, the bird's dissimilarity is present in its tail when contrasted with the kite and the airplane. Lastly, in the third task, the background of the bird image differs from the objects in other support images, i.e., the bamboo and bar.

In this paper, we propose a *Task-sensitive Discriminative Mutual Attention (TDMA)* network for few-shot classification, which adap-

* Corresponding Author. Emails: xubaogui2020@ruc.edu.cn (first author), bingsu@ruc.edu.cn (corresponding author)

tively enhances the discrimination ability of features with a new joint channel-spatial attention mechanism according to different tasks and image pairs. TDMA first generates a global discriminative task embedding based on linear discriminate analysis (LDA) to compute channel attentions, and the task embedding is used as the channel weight of the image. TDMA then generates spatial attention maps using a mutual attention module with the guidance of the task embedding to locate the discriminative parts in both images. The response at each position of one image is computed by taking all positions of the other image into consideration to tackle long-range dependencies. The generated channel-wise and spatial-wise attentions are used to enhance the features so that the discriminative foreground responses are strengthened and the irrelevant noisy information is filtered out. In this way, TDMA is able to improve the performance and does not bring too much computation and storage burden. Our contributions are threefold:

- We propose a novel LDA-based method which aims to search the most discriminative direction that the separability among classes of a task is maximized, referred to as "task embedding". We generate channel attentions from the task embedding to adaptively enhance channels that better discriminate different classes in different tasks.
- We propose a mutual attention module to adaptively locate the discriminative parts in both images. The mutual attention module establishes the correlations between the spatial positions of both images, where the global discriminative information of the task and all spatial positions of one image are incorporated to generate attention on the spatial positions of another. In this way, the mutual long-range dependencies and globally discriminative spatial responses are captured.
- We incorporate the task excitation network (TEN) and mutual attention module (MAM) into a joint discriminative model guided by task embeddings to learn both channel-wise attentions and spatial-wise attentions, and use the generated attentions to enhance the features. Experiments on three few-shot classification datasets show the effectiveness of our method.

2 Related Work

2.1 Few-shot Image Classification

The problem of few-shot learning is how to learn useful information for the current task from a small number of label samples in the new class. Depending on what meta-knowledge is learned, existing few-shot learning methods can be roughly divided into three categories: *optimization based*, *parameter-generating based*, and *metric-learning based*. Optimization-based methods [9, 25] aim to learn good initial model parameters so that the model can quickly adapt to new tasks through a limited number of gradient update steps. In their seminal work, MAML [9] presented a universal optimization algorithm designed to identify a specific set of model parameters that would generate significant performance enhancements on a novel task using a limited amount of training data and a small number of gradient steps. Reptile [25] introduced a modification to MAML that entailed removing the re-initialization step for every task, thereby rendering it more suitable for certain scenarios. This variant of MAML was demonstrated to be effective in the absence of task-specific re-initialization, as it converged towards a solution that was in proximity to the manifold of optimal solutions for each task. Parameter-generating based methods [23] learn an optimizer to predict meta-parameters. Metric-learning-based methods [36, 32, 33] embed input images into a common embedding space and learn a distance metric to distinguish samples from new classes using the nearest neighbor or nearest prototype-based classifier. An advantage of metric-learning methods is that only a simple feed-forward computation is required and no fine-tuning is needed on the target tasks. LDAM [40] dynamically sample local information and based on which to learn the position and channel-dependent relationship which has never been explored before, while DC [42] calibrate the distribution of few sample classes and use the expanded calibrated inputs to improve the classification effect, and it achieved good results.

In our proposed approach, we have integrated ProtoNet [32], a widely used metric-based framework to boost the generalization ability of the feature embeddings. Different from [32, 36, 33] where the same feature extractor is applied to all support and query images independently and hence all feature channels and spatial positions are treated equally, our method adaptively adjusts the features to enhance the discrimination ability when comparing different query and support samples. In [12], the cross attention network (CAN) also adapts the features of each pair of query and support images using the attention mechanism. Our method differs from CAN in three aspects. 1. CAN does not take into account the channel dimensions of feature maps, while our method generates both channel-wise and spatial-wise attentions to jointly identify discriminative features and locations. 2. CAN only exploits local point-wise relations to highlight the target object regions, while our method employs long-range dependencies using non-local operations to capture the relationship between support and query features interactively. 3. CAN produces features that are based on sample-specific, while our method is based on task-aware which can produce task-and-sample-specific features. [3] proposed a baseline++ method that aims at reducing intra-class variations by the cosine distances, the raise in inter-class separability of our task embedding is paralleled by the reduction of within-class variations. CAD [5] proposes a cross-attention module that computes attention between the query and support to generate the feature maps, however, it can only be conducted under the transductive setting since it uses all the query image information to enhance the representations of support images. Meanwhile, our method computes the channel and spatial attentions by computing the similarity scores and can work under the inductive setting.

Many few-shot learning methods that rely on prototypical networks have traditionally used L2 distance as a measure of the similarity between two images. However, recent research has introduced other distance metrics, such as Earth Mover's Distance (EMD) and Brownian Distance Covariance (BDC), which have led to promising results. Meta Navigator [44] presents a search space covering popular few-shot learning algorithms and a differentiable searching and decoding algorithm based on meta-learning for gradient-based optimization. It automates the selection of various few-shot learning designs and addresses the limitations of few-shot learning by finding good parameter adaptation policies for different stages in the network.

2.2 Attention Module

Attention mechanism has been widely used in many fields such as machine translation, natural language processing [7], and image classification [13]. In recent years, many attention-based methods have emerged in the field of computer vision, such as squeeze-and-excitation block [13], non-local attention block [37], and transformer [35]. Several recent studies have used attention mechanisms

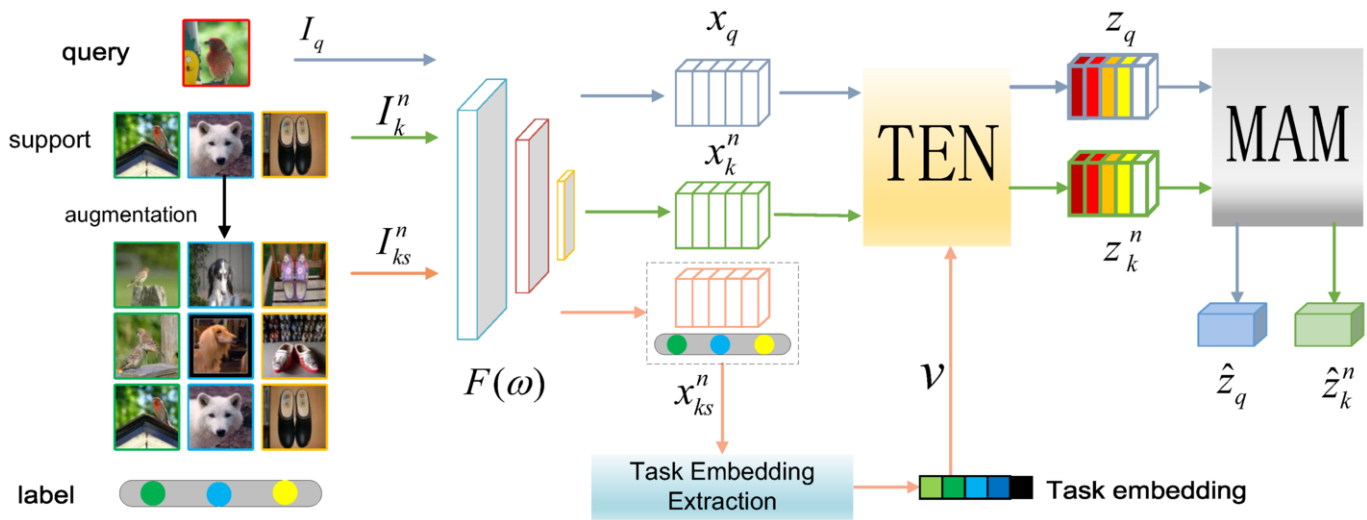

Figure 2: The diagram of the proposed TDMA for a 3-way 1-shot classification task with one query.

for few-shot learning [12, 43], they leverage self-attention and cross-attention to generate better feature representations. MELR [8] proposes a cross-episode attention module to take the attention between the cross-episode into consideration. In [43], the feature embeddings of instances are adapted to the target task using a transformer-based set-to-set function. In [46], the cross non-local neural network is proposed to capture the long-range dependency, which consists of several simplified non-local blocks. In this paper, we develop a meta-attention model to capture the task-aware discriminative information on channel dimension and learn the mutual relationships on spatial dimension, so that the model can quickly adapt to new tasks, even if there is little supervised information in new tasks. Specifically, the task embedding generates channel-wise attentions to select discriminative channel features, then our mutual attention module incorporates the task embedding to capture discriminative mutual attentions for each specific task.

3 Method

3.1 Problem Setting

Given the training set $\mathcal{D}^{train} = (I_i, y_i)$ with C base classes, where I_i is the i-th image sample and $y_i \in 1, \cdots, C$ is its corresponding label, few-shot learning aims at learning a model that can be quickly adapted to new tasks with a few labeled samples. For an N-way K-shot task in the testing phase, there are N classes that do not appear in the training set and only K labeled samples are available per class. The NK labeled samples form the support set of this task. The goal is to classify each unlabeled sample in a query set into one of the N classes. Usually, the tasks are constructed from testing set $\mathcal{D}^{test}$, where the classes in $\mathcal{D}^{test}$ are disjoint with those in $\mathcal{D}^{train}$. To construct a task, N classes are randomly sampled. The support set and query set are generated by sampling K and Q samples per class, respectively. The performance of the few-shot learning model is evaluated on these testing tasks.

3.2 Overview

The central concept of TDMA is depicted in Figure 2. Given a few-shot classification task, TDMA initially extracts a task embedding and then generates channel-wise attention to enhance the task-discriminative channels that are discriminative for the current task. For a pair of enhanced support and query feature maps, TDMA utilizes a task-sensitive module to create sample-pair-specific mutual attentions that pinpoint the discriminative regions within the feature maps, resulting in sample-pair-specific feature maps for the query and support. Finally, TDMA adopts the ProtoNet [32] to perform classification for the adapted support and query features.

3.3 Task Embedding Extraction

For an N-way K-shot classification task, the support set is denoted by $(\boldsymbol{I}_k^n, n = 1, \cdots, N, k = 1, \cdots, K)$. For the k-th support image $\boldsymbol{I}_k^n$ of the n-th class, we generate a number of S augmented images $\boldsymbol{I}_{ks}^n, s = 1, \cdots, S$ from it by RandAugment [6] and TrivialAugment [24]. $\boldsymbol{I}_k^n$ and all the augmentations $\boldsymbol{I}_{ks}^n, s = 1, \cdots, S$ are fed into a feature extractor backbone $F(\omega)$ to obtain the feature maps $\boldsymbol{x}_k^n = F(\boldsymbol{I}_k^n; \omega) \in \mathbb{R}^{C \times H \times W}$ and $\boldsymbol{x}_{ks}^n = F(\boldsymbol{I}_{ks}^n; \omega) \in \mathbb{R}^{C \times H \times W}, s = 1, \cdots, S$, where C, H and W are the number of channels, height and width of the feature map, respectively. We perform the global average pooling [19] on all the feature maps and obtain the corresponding feature vectors $\boldsymbol{x}_k^n$ and $\boldsymbol{x}_{ks}^n$, respectively.

For the n-th class, we calculate the mean of all feature vectors of all the support images and their augmentations:

$$\bar{\boldsymbol{x}^n} = \frac{1}{K(S+1)} \sum_{k=1}^{K} (\boldsymbol{x}_k^n + \sum_{s=1}^{S} \boldsymbol{x}_{ks}^n). \tag{1}$$

The within-class covariance of the n-th class can be obtained as:

$$\begin{aligned} \boldsymbol{\Gamma}^n = \frac{1}{K(S+1)} \sum_{k=1}^{K} ((\boldsymbol{x}_k^n - \bar{\boldsymbol{x}^n})(\boldsymbol{x}_k^n - \bar{\boldsymbol{x}^n})^T + \\ \sum_{s=1}^{S} (\boldsymbol{x}_{ks}^n - \bar{\boldsymbol{x}^n})(\boldsymbol{x}_{ks}^n - \bar{\boldsymbol{x}^n})^T). \end{aligned} \tag{2}$$

The overall intra-class scatter is the weighted average of covariances of all the support classes:

$$\boldsymbol{\Gamma}_w = \sum_{n=1}^{N} p^n \boldsymbol{\Gamma}^n, \tag{3}$$

where $p^n = \frac{1}{N}$ is the prior probability of n-th class in the task. The inter-class scatter can be calculated as follows:

$$\mathbf{\Gamma}_b = \sum_{n=1}^{N-1} \sum_{n'=n+1}^{N} (\bar{\boldsymbol{x}^n} - \bar{\boldsymbol{x}^{n'}})(\bar{\boldsymbol{x}^n} - \bar{\boldsymbol{x}^{n'}})^T. \tag{4}$$

$\mathbf{\Gamma}_w$ and $\mathbf{\Gamma}_b$ measure the intra-class compactness and the separability among different classes, respectively. We apply Linear Discriminative Analysis (LDA) [10, 27] to capture the most discriminative projection $\boldsymbol{v}$:

$$\boldsymbol{v} = \underset{\boldsymbol{w}}{argmax} \frac{tr(\boldsymbol{w}^T \mathbf{\Gamma}_b \boldsymbol{w})}{tr(\boldsymbol{w}^T \mathbf{\Gamma}_w \boldsymbol{w})}. \tag{5}$$

Here tr represents the trace of the matrix. The solution of Eq.(5) is the most dominant generalized eigenvector of the matrix $\mathbf{\Gamma}_w^{-1}\mathbf{\Gamma}_b$ w.r.t. the largest eigenvalue. By projecting all feature vectors of all classes with $\boldsymbol{v}$, the ratio of the inter-class scatter over the intra-class scatter is maximized, i.e., the separability of different classes is maximized. Therefore, $\boldsymbol{v}$ encodes the most discriminative information among classes in the task. We use $\boldsymbol{v}$ as the task embedding to guide the extraction of channel-wise and spatial attentions for comparing different query and support images.

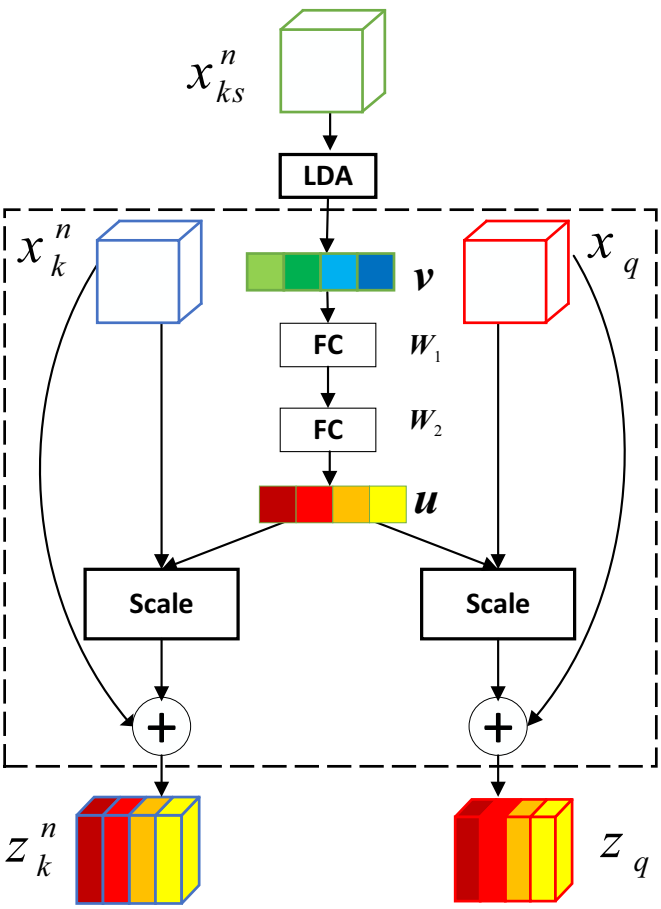

Figure 3: Task excitation network(TEN).

3.4 Task Excitation Network

The task excitation module is used to generate discriminative channel-wise attentions from the task embedding $\boldsymbol{v}$, which has a similar architecture to the squeeze and excitation (SE) module [13]. As shown in Figure 3, this module has three inputs: task embedding $\boldsymbol{v}$, query feature $\boldsymbol{x}_q$ and support feature $\boldsymbol{x}_k^n$. Specifically, it contains two fully connected (FC) layers each followed by a sigmoid activation. Different from the SE module, it takes the task embedding $\boldsymbol{v}$ as the input instead of the average pooling operation. The channel-wise attention $\boldsymbol{u}$ is calculated as follows:

$$\boldsymbol{u} = \sigma(\boldsymbol{W}\boldsymbol{v}) = \sigma(\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{v}), \tag{6}$$

where σ denotes the sigmoid function, $\boldsymbol{W}_1 \in \mathbb{R}^{C \times \frac{C}{r}}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{\frac{C}{r} \times C}$ are the weights of the two FC layers, and r is the scaling ratio, which is usually set as a constant. This equation is designed to fine-tune the task embedding into the channel attention.

The learned channel-wise attention $\boldsymbol{u}$ assigns larger weights to those discriminative channels for the task. $\boldsymbol{u}$ is applied to the feature maps of both support and query images. The enhanced feature map $\boldsymbol{z}_k^n$ for the support feature map $\boldsymbol{x}_k^n$ is calculated as:

$$\boldsymbol{z}_k^n = \boldsymbol{x}_k^n \otimes \boldsymbol{u} + \boldsymbol{x}_k^n, \tag{7}$$

where $\otimes$ is the channel-wise multiplication.

A query image $\boldsymbol{I}_q$ is first fed into the backbone $F(\omega)$ to obtain the query feature map $\boldsymbol{x}_q = F(\boldsymbol{I}_q; \omega) \in \mathbb{R}^{C \times H \times W}$, which is also enhanced by the channel-wise attention $\boldsymbol{u}$ as follows:

$$\boldsymbol{z}_q = \boldsymbol{x}_q \otimes \boldsymbol{u} + \boldsymbol{x}_q. \tag{8}$$

Then the output of the enhanced query and support image features, $\boldsymbol{x}_q$ and $\boldsymbol{x}_k^n$, are rendered more discriminative through channel weighting facilitated by the task embedding.

3.5 Mutual Attention Module

Locations of objects and discriminative parts differ in images and tasks. The relevant regions in an image are also different when compared to other images in different tasks. We design a task-aware mutual attention module to locate the most discriminative regions adaptively in each specific pair of support and query images for each specific task by establishing their mutual spatial-wise relationships. Specifically, mutual attention means the learned query attention to the support feature map and the learned support attention for the query feature map, thus different pairs of support and query images have different mutual attention. Besides, for the task-specific mutual attention calculation, the mutual attention module engages the task embedding with the support or query feature map to locate the discriminative part of the image. We design the task-sensitive mutual attention module to locate the discriminative regions adaptively in each specific pair of support and query images for each specific task by establishing their mutual spatial-wise relationships as shown in Figure 4.

Once guided by the task embedding, there is an expectation that the query and support images will engage in mutual learning. In light of the aforementioned, a mutual learning model was developed to facilitate bi-directional knowledge transfer between the query and support images, which can be formulated as follows:

$$\hat{\boldsymbol{z}}_q = softmax(\frac{W_\theta \boldsymbol{z}_k^n (W_\phi \boldsymbol{z}_q)^T}{\sqrt{d}}) \odot (W_g \boldsymbol{z}_q) + \boldsymbol{z}_q, \tag{9}$$

$$\hat{\boldsymbol{z}}_k^n = softmax(\frac{W_\theta \boldsymbol{z}_q (W_\phi \boldsymbol{z}_k^n)^T}{\sqrt{d}}) \odot (W_g \boldsymbol{z}_k^n) + \boldsymbol{z}_k^n, \tag{10}$$

where $\odot$ is the element-wise multiplication, and d is the dimension of the image feature. W_θ, W_ϕ, and W_g are weight matrices

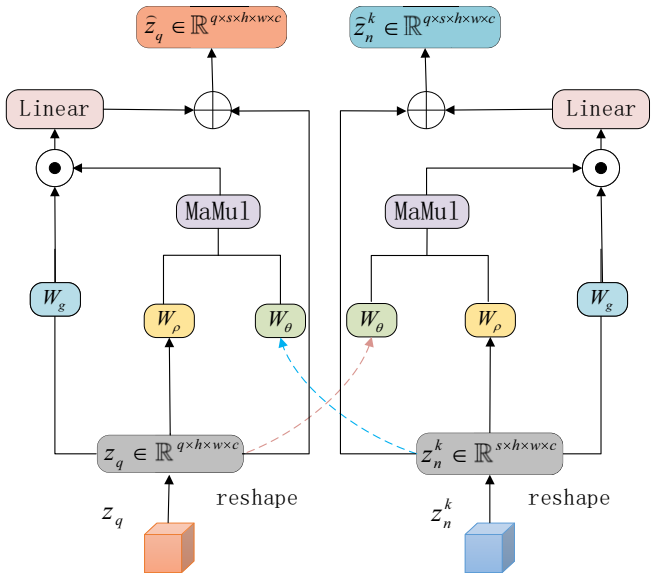

Figure 4: The task-sensitive mutual attention module(MAM).

to be learned. $\hat{\boldsymbol{z}}_q$ and $\hat{\boldsymbol{z}}_k^n$ are the output of the spatial-enhanced feature maps for the support and query feature map respectively. $W_\theta \boldsymbol{z}_k^n (W_\phi \boldsymbol{z}_q)^T$ and $W_\theta \boldsymbol{z}_q (W_\phi \boldsymbol{z}_k^n)^T$ are the similarity matrices between the query and support. For query features, the element-wise multiplication between $W_\theta \boldsymbol{z}_k^n (W_\phi \boldsymbol{z}_q)^T$ and $W_g \boldsymbol{z}_q$ means the features need to be enhanced for the query. Spatial attention is obtained through computing attention for all of the spatial points (hw) in the feature map.

The proposed mutual attention module captures the spatial-wise correspondences and inter-dependencies between two feature maps with the guidance of channel-wise attention. The learned discriminative attention can effectively locate the relevant areas that appear together between the support and query images, regardless of whether these highly relevant areas are associated with objects from seen classes or new classes, and hence can be generalized to new classes.

3.6 Classification and Model training

For a pair of support image $\boldsymbol{I}_k^n$ and a query image $\boldsymbol{I}_q$, the mutual attention module generates the re-weighted support feature map $\hat{\boldsymbol{z}}_k^n$ and query feature map $\hat{\boldsymbol{z}}_q$, respectively. The same query image has different attention maps when comparing different support images. We then employ the Euclidean distance $d(\cdot,\cdot)$ to measure the similarity between $\hat{\boldsymbol{z}}_q$ and $\hat{\boldsymbol{z}}_k^n$. In the training stage, we follow the metric-learning-based method [32] to calculate the prototype of the n-th class by averaging the feature maps of the support, and we use $\hat{\boldsymbol{z}}_\cdot^n$ to denote the prototype of the support image features. Then the prototype of and support feature map could be used to minimize the classification loss on the query set of the training episodes. The classification loss is defined as Eq.(11):

$$\mathcal{L} = \frac{1}{||Q||} \sum_{q=1}^{||Q||} -log \frac{exp(-d(\hat{\boldsymbol{z}}_q, \hat{\boldsymbol{z}}_\cdot^{y_q}))}{\sum_{n \in N} exp(-d(\hat{\boldsymbol{z}}_q, \hat{\boldsymbol{z}}_\cdot^{n}))}, \tag{11}$$

where $\hat{\boldsymbol{z}}_\cdot^n = \frac{1}{K} \sum_{k=1}^{K} \hat{\boldsymbol{z}}_k^n$, $\hat{\boldsymbol{z}}_\cdot^n$ is the prototype embedding of n-th class. $||Q||$ is the number of query images per episode. y_q is the corresponding label of the $\hat{\boldsymbol{z}}_q$.

Algorithm 1 A single episode update for training TDMA

Input: A N-way K-shot episode with support images $\boldsymbol{I}_k^n$ and query images $\boldsymbol{I}_q$
Initialize: feature extractor backbone $F(\omega)$, weights of convolutional layers W_θ, W_ϕ, W_g
for e in $\{E_1, \cdots, E_E\}$ **do**
  for i in $\{1, \cdots, N\}$ **do**
    1. Generate augmentations for support images $\boldsymbol{I}_{ks}^n$;
    2. Extract features $\boldsymbol{x}_{ks}^n$, $\boldsymbol{x}_k^n$, $\boldsymbol{x}_q$ for $\boldsymbol{I}_{ks}^n$, $\boldsymbol{I}_k^n$, $\boldsymbol{I}_q$ by $F(\omega)$, respectively;
    3. Calculate the task embedding $\boldsymbol{v}$ using $\boldsymbol{I}_{ks}^n$ by Eq.(1)-Eq.(5);
    4. Compute the enhanced support and query feature map $\boldsymbol{z}_k^n$ and $\boldsymbol{z}_q$ by the task embedding $\boldsymbol{v}$ using Eq.(6)-Eq.(8);
    4. Calculate the mutual attention feature maps $\hat{\boldsymbol{z}}_k^n$, $\hat{\boldsymbol{z}}_q$ by Eq.(9)-Eq.(10)
    5. Calculate $\mathcal{L}_\mathcal{T}$ using Eq.(11)
    6. Update $F(\omega)$, W_θ, W_ϕ, W_g with respect to $\mathcal{L}_\mathcal{T}$
  end for
end for

For the testing stage, The predicted label for the query image x_q is calculated as follows:

$$\hat{y}_q = \arg\max_{y \in N} \frac{exp((-d(\hat{\boldsymbol{z}}_q, \hat{\boldsymbol{z}}_\cdot^{y}))}{\sum_{n \in N} exp(-d(\hat{\boldsymbol{z}}_q, \hat{\boldsymbol{z}}_\cdot^{n}))}. \tag{12}$$

3.7 Overall TDMA Algorithm

To ensure reproducibility, we provide a complete outline of the algorithm for Few-Shot Learning (FSL) with our TDMA, as depicted in Algorithm 1. With the learned model, we can carry out inference on the test episodes.

Complexity Analysis. The complexities of the task embedding extraction, the task excitation network, and the mutual attention module are $O(C^3)$, $O(C^2)$, and $O(H^2W^2C)$, respectively.

4 Experiments

Datasets. We evaluate the proposed method on three widely-used datasets, including miniImageNet [36], tieredImageNet [30] and FC100 [44]. The first two datasets are subsets of ILSVRC-12 [31], in which the image size is 84×84. MiniImageNet contains 100 classes, which are divided into a training set of 64 classes, a validation set of 16 classes, and a test set of 20 classes. TieredImageNet contains 608 categories(779,165 images) and is divided into a training set of 351 classes, a validation set of 97 classes, and a test set of 160 classes. FC100 is a few-shot classification dataset consisting of 100 distinct object classes derived from CIFAR100 [15], with each class being represented by a collection of 32×32 color images, amounting to 600 samples per class. Following the prior work [44] for task splits, we partitioned the dataset into training, validation, and testing subsets, containing 60, 20, and 20 classes, respectively.

Evaluation Metrics. In line with previous works, we take the 5-way 1-shot/5-shot FSL evaluation setting. The evaluation process random sample 5 classes for each episode with 1-shot/5-shot and 15 query images per class, thus every test episode has 75 query images and 5 or 25 support image(s). The results are reported on average 5-way classification accuracy (%, top-1) over randomly sampled 2,000 test episodes as well as the 95% confidence interval. As the task excitation network and mutual attention module cope with each sample

independently, our TDMA is evaluated under a strict inductive setting.

Feature Extractor. To ensure a fair comparison with prior literature, our proposed TDMA method employs the commonly used ResNet-12 [11] as the underlying feature extractor. To expedite the training process, we initialize the feature extractor by pretraining it on the training partition of each dataset following the common practice, which has been observed in numerous prior studies [43, 45].

Table 1: The mean accuracy (%, top-1) results of the standard FSL along with a 95% confidence interval on the miniImageNet. The top two results are presented in bold and underlined format respectively.

Method	Backbone	5-way 1-shot	5-way 5-shot
MAML [9]	Conv4	48.70 ± 1.84	63.11 ± 0.92
MatchingNet [36]	Conv4	43.56 ± 0.84	55.31 ± 0.73
ProtoNet [32]	Conv4	49.42 ± 0.78	68.20 ± 0.66
ECSF [29]	Conv4	49.07 ± 0.43	65.73 ± 0.36
BOIL [26]	Conv4	49.61 ± 0.16	66.45 ± 0.37
IEPT [45]	Conv4	56.26 ± 0.45	73.91 ± 0.34
LDGP [38]	Conv4	56.32 ± 0.28	72.64 ± 0.26
CAN [12]	ResNet-12	62.64 ± 0.66	78.83 ± 0.45
FEAT [43]	ResNet-12	66.78 ± 0.20	82.05 ± 0.14
infoPatch [20]	ResNet-12	67.04 ± 0.63	83.63 ± 0.29
CNL [46]	ResNet-12	67.96 ± 0.98	83.36 ± 0.51
BML [48]	ResNet-12	67.67 ± 0.45	82.44 ± 0.29
ConstellationNet [41]	ResNet-12	64.89 ± 0.23	79.95 ± 0.17
IEPT [45]	ResNet-12	67.05 ± 0.44	82.90 ± 0.30
DMF [39]	ResNet-12	67.76 ± 0.46	82.71 ± 0.31
RENET [14]	ResNet-12	67.60 ± 0.44	82.58 ± 0.30
SetFeat [2]	ResNet-12	68.32 ± 0.62	82.71 ±0.46
Meta-Baseline [4]	ResNet-12	63.17 ± 0.23	79.26 ± 0.17
NCA NC [16]	ResNet-12	62.55 ± 0.12	78.27 ± 0.09
POODLE [17]	ResNet-12	67.80	83.50
PAL [22]	ResNet-12	69.37 ± 0.64	84.40 ± 0.44
TDMA(ours)	ResNet-12	**70.42 ± 0.46**	**84.52 ± 0.30**

Table 2: The mean accuracy (%, top-1) results of the standard FSL with a 95% confidence interval on the tieredImageNet. The top two results are presented in bold and underlined format, respectively.

Method	Backbone	5-way 1-shot	5-way 5-shot
MAML [9]	Conv4	51.67 ± 1.81	70.30 ± 1.75
ProtoNet [32]	Conv4	53.31 ± 0.89	72.69 ± 0.74
ECSF [29]	Conv4	48.19 ± 0.43	65.50 ± 0.39
IEPT [45]	Conv4	58.25 ± 0.48	75.63 ± 0.0.46
BOIL [26]	Conv4	48.58 ± 0.27	69.37 ± 0.12
LDGP [38]	Conv4	58.43 ± 0.38	76.17 ± 0.34
CAN [12]	ResNet-12	66.22 ± 0.75	82.79 ± 0.48
FEAT [43]	ResNet-12	70.80 ± 0.23	84.79 ± 0.16
Rethink-Distill [34]	ResNet-12	71.52 ± 0.69	86.03 ± 0.49
infoPatch [20]	ResNet-12	71.51 ± 0.52	85.44 ± 0.35
DMF [39]	ResNet-12	71.89 ± 0.52	82.71 ± 0.31
RENET [14]	ResNet-12	71.61 ± 0.51	85.28 ± 0.35
BML [48]	ResNet-12	68.99 ± 0.50	85.49± 0.34
Meta-Baseline [4]	ResNet-12	68.62 ± 0.27	83.74 ± 0.18
IEPT [45]	ResNet-12	72.24 ± 0.50	86.73 ±0.34
NCA NC [16]	ResNet-12	68.35 ± 0.13	83.20 ± 0.10
POODLE [17]	ResNet-12	70.42	85.26
PAL [22]	ResNet-12	72.25 ± 0.72	**86.95 ± 0.47**
TDMA(ours)	ResNet-12	**72.57 ± 0.51**	86.02 ± 0.16

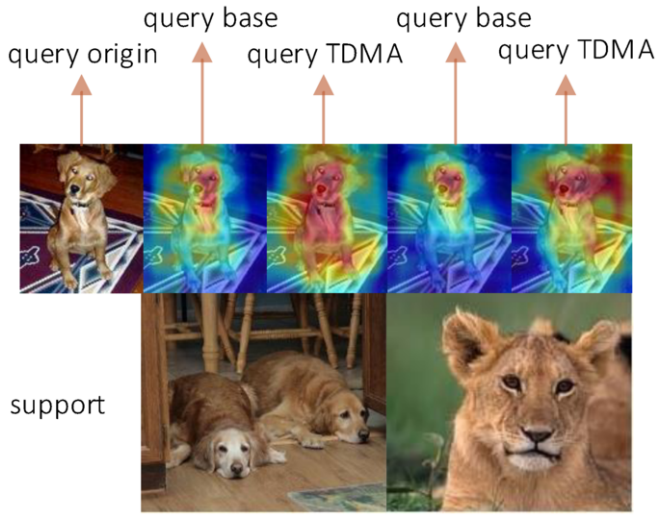

Figure 5: Ablation result of class activation mapping visualization on a 5-way 1-shot task with one query, we only show 2-way of it.

4.1 Main Results

As shown in Table 1 and 2, TDMA remarkably outperforms the state-of-the-art inductive few-shot learning methods on miniImageNet and obtains competitive results on tieredImageNet. Table 3 shows that we also achieve comparable results on FC100. In experiments conducted on miniImageNet and tieredImageNet, TDMA exhibits a considerable performance advantage over two other attention-based methodologies, CAN and FEAT. We would compare the attention mechanism employed in these different methods. The attentions learned by FEAT mainly focus on the support images of all classes; CAN generates cross-attentions between the support and query images by emphasizing the similar parts across spatial positions but does not take the overall task into consideration, causing the risk that common parts may not be related to the foreground areas and the task. Differently, Conversely, our proposed TDMA method leverages task embeddings to identify distinctive spatial regions, thereby yielding superior performance in comparison to CAN and FEAT.

Table 3: The mean accuracy (%, top-1) results of the standard FSL with a 95% confidence interval on the FC100. The top two results are presented in bold and underlined format, respectively.

Method	Backbone	5-way 1-shot	5-way 5-shot
MAML [9]	Conv4	38.1 ± 1.7	50.4 ± 1.0
ProtoNet [32]	Conv4	41.54 ± 0.76	57.08± 0.76
E^3BM [21]	ResNet-12	43.2 ± 0.30	60.2 ± 0.30
Centroid [1]	ResNet-12	45.83 ± 0.48	59.74± 0.56
Rethink-Distill [34]	ResNet-12	44.6 ± 0.70	60.9 ± 0.60
ConstellationNet [41]	ResNet-12	43.8 ± 0.20	59.7 ± 0.20
infoPatch [20]	ResNet-12	43.8 ± 0.40	58.0 ± 0.4
MixtFSL [48]	ResNet-12	44.89 ± 0.63	60.7± 0.60
PAL [22]	ResNet-12	**47.2 ± 0.6**	**64.0± 0.60**
TDMA(ours)	ResNet-12	45.09 ± 0.41	60.82 ± 0.41

Table 4: Ablation study of the components in on miniImageNet.

	TDMA	*w/o* MAM	*w/o* TEN
5-way 1-shot	70.42±0.46	69.58±0.46	69.95 ± 0.46

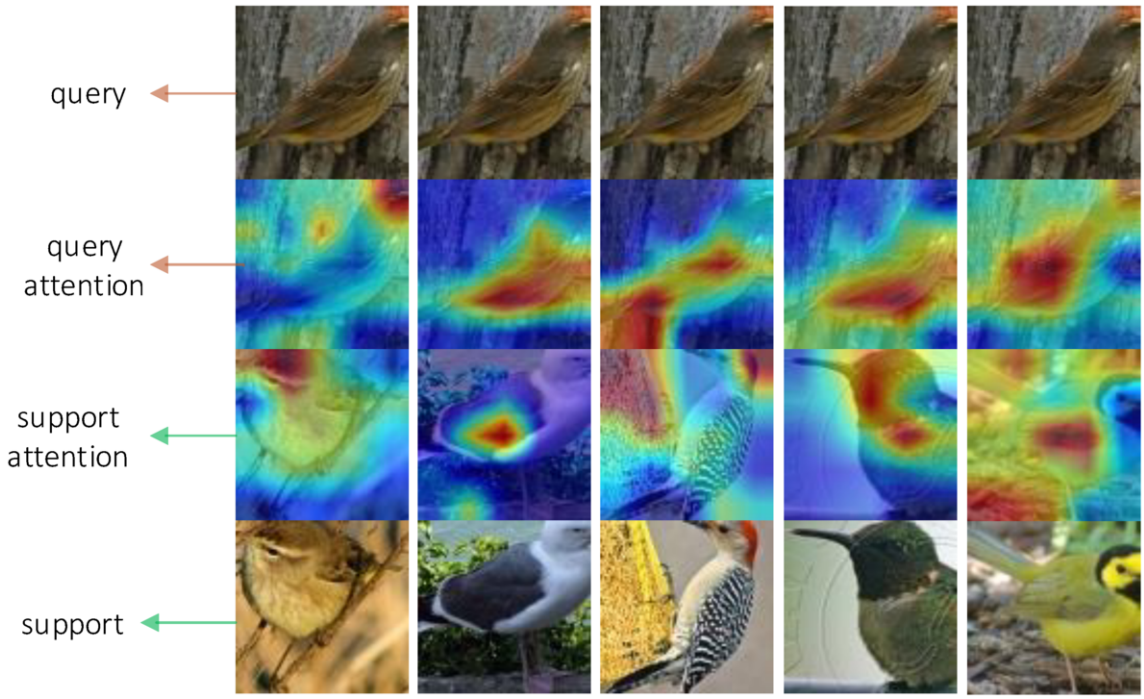

Figure 6: The first row of the map corresponds to a single query image, with the final row representing the support images. The intermediate rows depict the heat map for the query and support, respectively. Specifically, the leftmost support image in the fourth row belongs to the same class as the query image, with the heat map primarily highlighting the bird's head. In contrast, for other image pairs, the attention map is more dispersed, with attention being directed towards diverse image regions that aid in differentiation.

Ablation study. On top of the base method, TDMA contains two new modules: TEN and MAM. Table 4 shows that TEN and MAM can improve the model's discrimination ability. Figure 5 illustrates the activation maps [47] of the base method and TDMA. "query origin" and "support" are the query and support images respectively, "query base" means the attention maps of the query generated by the base method, and "query TDMA" means the attention maps of query calculated by our TDMA. It can be seen that the activation maps of the query image generated by the base model are quite similar even when the support images are different. By contrast, the attention learned by TDMA focuses on dog ears in the query image when the support image is a lion picture. That is to say, TEN and MAM are capable of extracting task-specific features and spatial regions that possess discriminative qualities, facilitating the classifier's ability to distinguish between different classes.

4.2 Visualization Analysis

TDMA produces task embedding to guide the learning of spatial-wise mutual attention, aiding in identifying the distinguishing features within images. Inspired by Class Activation Map (CAM) [47], we create an attention map to establish associations between query and support images, with a one-to-one correspondence between each pair. Figure 6 depicts the heat map, where high-intensity regions correspond to discriminative regions, except for the first support image, which emphasizes similarity with the query image.

5 Conclusion

This paper introduces TDMA, a novel approach for few-shot image classification, which involves extracting a task embedding to capture the most discriminative direction to differentiate between various support classes. Channel-wise attentions are generated based on this task embedding to emphasize relevant and informative features for each task. Additionally, TDMA employs a mutual attention module to refine selected features in query and support image pairs, thus enhancing model discriminative ability. Extensive evaluations on benchmark datasets demonstrate that TDMA achieves competitive performance on these datasets over existing state-of-the-art methods. Overall, our results show the effectiveness of our proposed method in addressing the few-shot image classification problem.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China No. 61976206 and No. 61832017, Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098, Beijing Academy of Artificial Intelligence (BAAI), the Fundamental Research Funds for the Central Universities, the Research Funds of Renmin University of China 21XNLG05, and Public Computing Cloud, Renmin University of China.

References

[1] Arman Afrasiyabi, Jean-François Lalonde, and Christian Gagné, 'Associative alignment for few-shot image classification', in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pp. 18–35. Springer, (2020).

[2] Arman Afrasiyabi, Hugo Larochelle, Jean-François Lalonde, and Christian Gagné, 'Matching feature sets for few-shot image classification', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9014–9024, (2022).

[3] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang, 'A closer look at few-shot classification', *arXiv preprint arXiv:1904.04232*, (2019).

[4] Yinbo Chen, Zhuang Liu, Huijuan Xu, Trevor Darrell, and Xiaolong Wang, 'Meta-baseline: Exploring simple meta-learning for few-shot learning', in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9062–9071, (2021).

[5] Philip Chikontwe, Soopil Kim, and Sang Hyun Park, 'Cad: Co-adapting discriminative features for improved few-shot classification', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14554–14563, (2022).

[6] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le, 'Randaugment: Practical automated data augmentation with a reduced search space', (2019).

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 'Bert: Pre-training of deep bidirectional transformers for language understanding', *arXiv preprint arXiv:1810.04805*, (2018).

[8] Nanyi Fei, Zhiwu Lu, Tao Xiang, and Songfang Huang, 'Melr: Meta-learning via modeling episode-level relationships for few-shot learning', in *Proc. Int. Conf. Learn. Represent.*, pp. 1–20, (2021).

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine, 'Model-agnostic meta-learning for fast adaptation of deep networks', in *International Conference on Machine Learning*, pp. 1126–1135. PMLR, (2017).

[10] Ronald A Fisher, 'The use of multiple measurements in taxonomic problems', *Annals of eugenics*, **7**(2), 179–188, (1936).

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, 'Deep residual learning for image recognition', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, (2016).

[12] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen, 'Cross attention network for few-shot classification', in *NeurIPS*, (2019).

[13] Jie Hu, Li Shen, and Gang Sun, 'Squeeze-and-excitation networks', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, (2018).

[14] Dahyun Kang, Heeseung Kwon, Juhong Min, and Minsu Cho, 'Relational embedding for few-shot classification', in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8822–8833, (2021).

[15] Alex Krizhevsky, Geoffrey Hinton, et al., 'Learning multiple layers of features from tiny images', (2009).

[16] Steinar Laenen and Luca Bertinetto, 'On episodes, prototypical networks, and few-shot learning', *Advances in Neural Information Processing Systems*, **34**, 24581–24592, (2021).

[17] Duong Le, Khoi Duc Nguyen, Khoi Nguyen, Quoc-Huy Tran, Rang Nguyen, and Binh-Son Hua, 'Poodle: Improving few-shot learning via penalizing out-of-distribution samples', *Advances in Neural Information Processing Systems*, **34**, 23942–23955, (2021).

[18] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto, 'Meta-learning with differentiable convex optimization', in *CVPR*, (2019).

[19] Min Lin, Qiang Chen, and Shuicheng Yan, 'Network in network', in *ICLR*, (2014).

[20] Chen Liu, Yanwei Fu, Chengming Xu, Siqian Yang, Jilin Li, Chengjie Wang, and Li Zhang, 'Learning a few-shot embedding model with contrastive learning', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8635–8643, (2021).

[21] Yaoyao Liu, Bernt Schiele, and Qianru Sun, 'An ensemble of epoch-wise empirical bayes for few-shot learning', in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pp. 404–421. Springer, (2020).

[22] Jiawei Ma, Hanchen Xie, Guangxing Han, Shih Fu Chang, Aram Galstyan, and Wael Abd-Almageed, 'Partner-assisted learning for few-shot image classification', (2021).

[23] Tsendsuren Munkhdalai and Hong Yu, 'Meta networks', in *International Conference on Machine Learning*, pp. 2554–2563. PMLR, (2017).

[24] Samuel G. Müller and Frank Hutter, 'Trivialaugment: Tuning-free yet state-of-the-art data augmentation', (2021).

[25] Alex Nichol, Joshua Achiam, and John Schulman, 'On first-order meta-learning algorithms', *arXiv preprint arXiv:1803.02999*, (2018).

[26] Jaehoon Oh, Hyungjun Yoo, ChangHwan Kim, and Se-Young Yun, 'Boil: Towards representation change for few-shot learning', *arXiv preprint arXiv:2008.08882*, (2020).

[27] C Radhakrishna Rao, 'The utilization of multiple measurements in problems of biological classification', *Journal of the Royal Statistical Society: Series B (Methodological)*, **10**(2), 159–193, (1948).

[28] Sachin Ravi and Hugo Larochelle, 'Optimization as a model for few-shot learning', in *ICLR*, (2016).

[29] Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto, 'Few-shot learning with embedded class models and shot-free meta training', in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 331–339, (2019).

[30] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel, 'Meta-learning for semi-supervised few-shot classification', *arXiv preprint arXiv:1803.00676*, (2018).

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, 'Imagenet large scale visual recognition challenge', *International Journal of Computer Vision*, **115**(3), 211–252, (2015).

[32] Jake Snell, Kevin Swersky, and Richard Zemel, 'Prototypical networks for few-shot learning', in *Advances in Neural Information Processing Systems*, (2017).

[33] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales, 'Learning to compare: Relation network for few-shot learning', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018).

[34] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola, 'Rethinking few-shot image classification: a good embedding is all you need?', in *European Conference on Computer Vision*, pp. 266–282. Springer, (2020).

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, 'Attention is all you need', *arXiv preprint arXiv:1706.03762*, (2017).

[36] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra, 'Matching networks for one shot learning', *arXiv preprint arXiv:1606.04080*, (2016).

[37] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, 'Non-local neural networks', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7794–7803, (2018).

[38] Ze Wang, Zichen Miao, Xiantong Zhen, and Qiang Qiu, 'Learning to learn dense gaussian processes for few-shot learning', *Advances in Neural Information Processing Systems*, **34**, 13230–13241, (2021).

[39] Chengming Xu, Yanwei Fu, Chen Liu, Chengjie Wang, Jilin Li, Feiyue Huang, Li Zhang, and Xiangyang Xue, 'Learning dynamic alignment via meta-filter for few-shot learning', pp. 5182–5191, (2021).

[40] Chengming Xu, Chen Liu, Li Zhang, Chengjie Wang, Jilin Li, Feiyue Huang, Xiangyang Xue, and Yanwei Fu, 'Learning dynamic alignment via meta-filter for few-shot learning', *arXiv preprint arXiv:2103.13582*, (2021).

[41] Weijian Xu, Yifan Xu, Huaijin Wang, and Zhuowen Tu, 'Attentional constellation nets for few-shot learning', in *International Conference on Learning Representations*, (2021).

[42] Shuo Yang, Lu Liu, and Min Xu, 'Free lunch for few-shot learning: Distribution calibration', *arXiv preprint arXiv:2101.06395*, (2021).

[43] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha, 'Few-shot learning via embedding adaptation with set-to-set functions', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8808–8817, (2020).

[44] Chi Zhang, Henghui Ding, Guosheng Lin, Ruibo Li, Changhu Wang, and Chunhua Shen, 'Meta navigator: Search for a good adaptation policy for few-shot learning', in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9435–9444, (2021).

[45] Manli Zhang, Jianhong Zhang, Zhiwu Lu, Tao Xiang, and Mingyu Ding, 'Iept: Instance-level and episode-level pretext tasks for few-shot learning', in *Proc. Int. Conf. Learn. Represent.*, pp. 1–16, (2021).

[46] Jiabao Zhao, Yifan Yang, Xin Lin, Jing Yang, and Liang He, 'Looking wider for better adaptive representation in few-shot learning', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10981–10989, (2021).

[47] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, 'Learning deep features for discriminative localization', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, (2016).

[48] Ziqi Zhou, Xi Qiu, Jiangtao Xie, Jianan Wu, and Chi Zhang, 'Binocular mutual learning for improving few-shot classification', in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8402–8411, (2021).